REJECTION PHENOMENA IN INTER-SIGNAL VOICE TRANSPLANTATIONS

Werner Verhelst and Henk Brouckxon

Vrije Universiteit Brussel, dept. ETRO-DSSP
Pleinlaan 2, B-1050 Brussels, Belgium
{wverhels, hbrouckx}@etro.vub.ac.be

ABSTRACT

The voice modification system studied in this paper allows to *transplant* selected voice characteristics (pitch, loudness, timing, and/or timbre) from a given *donor* utterance onto another (*patient*) utterance. Several potential applications exist for such a system, depending on the voice characteristics that are transplanted. The examples studied here are lip synchronization, voice dubbing and karaoke. Using overlap-add techniques for the actual speech modification, high quality results have been obtained that convincingly illustrate the potential of voice transplantation. In order to make voice transplantations widely applicable, their robustness should be further improved in order to avoid rejection of the transplanted characteristics, as discussed in some detail in the paper.

1. INTRODUCTION

The most important perceptual aspects of the human voice, like all audio signals, are pitch, loudness, timbre and their time evolution. In speech models, these characteristics are usually approximated as being independent of one another and as being determined by the acoustic signal's fundamental frequency f_0, amplitude, spectral envelope and time variation, respectively.

In this paper we discuss a voice modification system that allows speech to be generated from a mixture of acoustic parameter contours taken from different utterances of the same sentence material. Thus, for example, a hybrid utterance could be produced that has the same acoustic parameters and timing as an utterance $U2$, except for the pitch contour, which corresponds to the contour of another utterance $U1$. We have termed such systems *voice transplantation systems* [1] and we will discuss voice transplantation systems for applications such as lip synchronization, voice dubbing and karaoke in this paper. Section 2 presents the general architecture of the voice transplantation system and demonstrates that the well-known PSOLA algorithm [2] can be considered as a special type of pitch excited vocoder. Section 3 describes how the musical expression can be transplanted from one interpretation of a song to another, and introduces the notion of rejection phenomena that can cause perceived distortions. In section 4 several of these phenomena are identified in the context of a lip-synchronization application and a remedy in the form of an interactive editing procedure is proposed. It is concluded in section 5 that experiments have successfully illustrated the potential of high quality voice transplantations accross recordings of the human voice and that improving the robustness of transplantation systems could open the way for many exciting applications.

Support from the Flemish Community through grants from IWT and FWO is gratefully acknowledged.

2. VOICE TRANSPLANTATION SYSTEMS

2.1. General architecture

Over the past several years overlap-add (OLA) techniques have been proposed that allow high quality prosodic modification of speech [3]. In the analysis phase these algorithms represent f_0 and amplitude information of the input speech as a function of time, either explicitly or implicitly, while the spectral information is always represented implicitly by a sequence of short segments from the original signal.

The transplantation systems discussed in this paper rely on OLA techniques for high quality modification of voice characteristics and on dynamic time-warping (DTW) for proper time-alignment of voice characteristics that have been extracted from different utterances. The concept is illustrated in Fig. 1. A same text is read, spoken, or sung by a same or by different persons, resulting in two utterances $U1$ and $U2$. The acoustic parameter contours obtained after analysis of $U1$ are time-scaled such that the scaled contours fit the timing of $U2$. The appropriate time-scaling function is obtained using DTW, a technique that is well known from speech recognition [4]. Finally, a new utterance Ux is constructed after having selected for each acoustic parameter contour whether the time-scaled version of $U1$ or the version of $U2$ is to be used. This way, one can *transplant* selected speech characteristics from one utterance to another. Note that for each pair of utterances one can choose which utterance corresponds to $U1$ and which to $U2$. Thus, there is no loss of generality in choosing that the timing of Ux always corresponds to the timing of $U2$.

2.2. Dynamic time warping

A conventional DTW procedure can be applied to obtain the timing relationship between the two input utterances. First a matrix is constructed whose elements $d(j,i), j = 1 \ldots J, i = 1 \ldots I$ are the spectral distances between frames j of $U2$ and frames i of $U1$. J and I represent the number of frames in the respective signals $U2$ and $U1$. The time-warping path is obtained as the path (j_k, i_k) that minimizes the accumulated distance D

$$D = \sum_{k=1}^{N} d(j_k, i_k)$$

subject to the constraints $(j_1, i_1) = (1,1); (j_N, i_N) = (J, I)$ and $(j_{k-1}, i_{k-1}) \in \{(j_k - 1, i_k), (j_k - 1, i_k - 1), (j_k, i_k - 1)\}$

Dynamic Time Warping has been extensively studied for speech recognition and more sophisticated forms for the functional D and for the constraints have been proposed which improve recognition scores in a number of systems [4]. However, in [1] we could

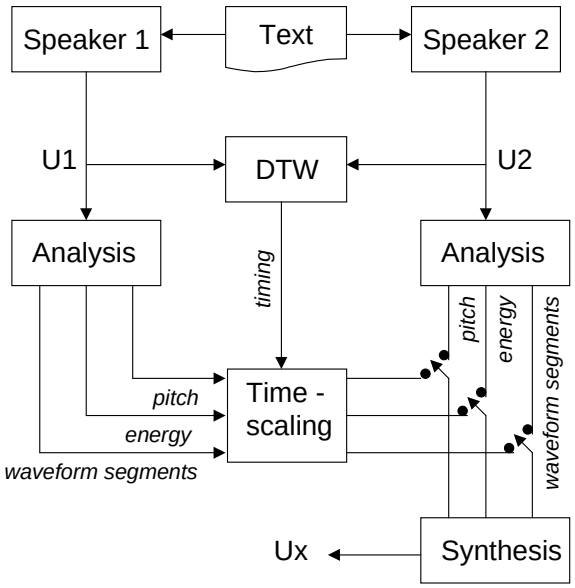

Figure 1: *Illustration of the voice transplantation concept. Two input utterances that correspond to a same text material are analyzed. The resulting parameter contours of one of them ($U1$) are time alligned to the other ($U2$). Each of the required resynthesis parameters can be chosen to originate from either $U1$ or $U2$. The result Ux has the same time structure as $U2$, and its pitch, loudness and timbre can each be chosen to originate from either $U1$ or $U2$.*

not clearly find them advantageous as far as the accuracy of the time-warping path was concerned (they often introduced inaccuracies when differences between acoustic realizations occurred, e.g., when one of the utterances contained a breathing pause). Therefore we opted to use the basic version of DTW described here.

2.3. Voice transplantation with PSOLA

PSOLA [2], [3] can be interpreted as a specific pitch-excited time varying linear system [5], as illustrated in fig. 2. The input sequence $i(n)$ to the system is constructed as an impulse train with unit impulses located at the analysis pitch marks:

$$i(n) = \sum_{k=-\infty}^{+\infty} \delta(n - p_a(k))$$

The sequence of analysis pitch marks $p_a(k)$ typically contains the sample indices of the zero-crossings at the beginning of consecutive pitch periods of the input sequence $x(n)$. The impulse response at time instants $p_a(k)$ is obtained by a simple windowing procedure applied to the input speech:

$$h(n, p_a(k)) = x(n).w(n - p_a(k))$$

where $w(n - p_a(k))$ is a two pitch period hanning-type window centered at $p_a(k)$.

In this way, $x(n)$ is analyzed to obtain pitch information $p_a(k)$ and synthesis filter impulse responses $h(n, p_a(k))$, which are also

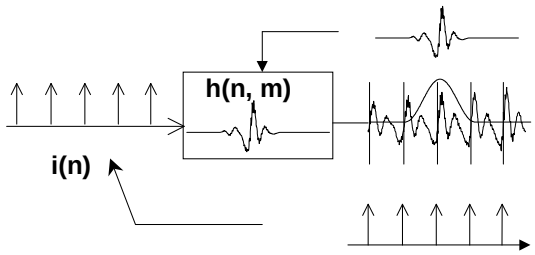

Figure 2: *Illustration of the pitch-excited system formulation for PSOLA*

the parameters used in traditional pitch excited vocoder schemes. Also, the modification and synthesis is similar to those of standard approaches like LPC vocoders (only here the synthesis filters are FIR filters defined at non-uniformly distributed sampling instants $p_a(k)$). Thus,

$$y(n) = \sum_{k=-\infty}^{+\infty} i(k)h(n, k)$$

where $i(k)$ and $h(n, k)$ represent the excitation signal and the synthesis filter parameters that are obtained by modifying the analysis parameters.

For pitch modification one generates a train of impulses spaced according to the desired pitch

$$i(k) = \sum_{l=-\infty}^{+\infty} \delta(k - p_s(l))$$

and the synthetic speech is then simply constructed as

$$y(n) = \sum_{k=-\infty}^{+\infty} h(n, p_s(k))$$

Impulse responses at times $p_s(k)$ are obtained by interpolation between the impulse responses that are available from the analysis: $h(n, m) = h(n, p_a(argmin_k|m - p_a(k)|))$ in the case of zero-order interpolation.

In a similar way, time scaling can be achieved by appropriately scaling the parameter tracks:

$$\begin{aligned} h_s(n, m) &= h_a(n, \tau^{-1}(m)) \\ p_s(k + 1) &= p_s(k) + p_a(T_{sa}(k) + 1) - p_a(T_{sa}(k)) \\ T_{sa}(k) &= argmin_l|\tau^{-1}(p_s(k)) - p_a(l)| \end{aligned}$$

In order to modify the energy contour of an utterance it suffices to apply appropriate gain factors to the successive synthesis impulse responses.

The transplantation of voice characteristics can be achieved by selecting the input utterance ($U1$ or $U2$) whose timbre corresponds to the desired output timbre and modifying the pitch, timing and gain of that utterance as desired. Note that all the required modifications can be performed in one step with PSOLA.

3. PROSODY TRANSPLANTATION AND KARAOKE

We initially used the voice transplantation system for prosody transplantation [1]. Here, the prosodic parameters (pitch, timing and loudness) of Ux are chosen to be those of $U2$, while the timbre of Ux is that of $U1$. As expected, with PSOLA the quality of the output after prosodic transplantation retained much of the naturalness and acoustic detail of the original input utterances.

However, informal experiments revealed that problems (*rejection phenomena*) could occur locally when the phonetic realizations of donor and patient utterances $U2$ and $U1$ differed. In Dutch, this situation frequently occurs for phonemes with a large number of allophones such as some fricatives, which can be produced with or without voicing, or such as [r], which can be dental or velar and may sound like a trill, a fricative or a glide. In such cases distortions can occur which are usually perceived as a segregated sound source, separated from the actual speech. For best results, it was found advisable to operate transplantations with patient/donor pairs that are phonetically as close as possible.

Additionally, care should be taken that PSOLA pitch marks are placed according to a strategy that remains consistent within and across both input utterances. Otherwise, a slight pitch jitter might be perceived. Picking zero-crossing locations for the pitch marks is usually easy and adequate for intra-speaker transplantations, but may be less adequate for inter-speaker transplantations.

Recently, transplantation of pitch, timing and energy between sung voices was studied by Kim Lau in an effort to create an off-line "super karaoke" that corrects the musical expression of the user's voice to match that of the original artist [6], [7]. This turned out to be an extremely challenging task: a.o., the wide range of pitches used in singing voices makes the analysis difficult and sometimes stresses the capabilities of OLA based voice modification algorithms. Although Kim concluded that his system has yet to attain a desirable level of robustness, it did successfully demonstrate the idea of "hybridizing" vocal performances and "super karaoke".

4. LIP SYNCHRONIZATION AND VOICE DUBBING

For this kind of applications, the transplantation system should produce a Ux with all the acoustic parameters from $U1$ and the timing of $U2$. The result should thus ideally correspond to a time-scaled version of $U1$ that is synchronized to $U2$. This would allow for automatically correcting lip-sync errors in postsynchronization work [8] as the timing from the reference track on the video or film could be transplanted on the replacement track recorded in the studio. Additionally, it could also be used to create artificial chorus effects.

4.1. Implementation

We used the time-scaling algorithm WSOLA [9] to produce the time-scaled version of the original in accordance with the time-warping path. Compared to PSOLA, WSOLA has the advantage that it operates pitch asynchronously and does not rely on pitch marks, which makes it easier to operate and more robust. Since it does not allow for pitch changes, however, WSOLA could not be used for the prosodic modification and karaoke applications.

4.2. Diagnostic evaluation

The time scaling accuracy of WSOLA appeared to be sufficiently high for this application. WSOLA operates with a timing tolerance of $[-\Delta_{max} \ldots + \Delta_{max}]$ in order to ensure pitch continuity in the time-scaled signal [9]. With a tolerance $\Delta_{max} = 7ms$, pitch continuity could be ensured without introducing noticeable time-misalignments.

Much like in the prosody transplantation and karaoke applications, the system demonstrated good performance, but lacked robustness and regularly severe distortions occurred. These distortions could often be traced to some event in the time-warping path, but could not always be considered to be due to an error in the path.

Generally, the results with the current system were good when the timing reference and the original were acoustically and phonetically sufficiently close since distortions could often be attributed to different types of acoustic-phonetic differences:

Impossible insertions. When a phone is pronounced carefully in the timing reference, but is absent or heavily coarticulated and short-lived in the original, a correct synchronization would require that a few transient frames of the original be heavily time-stretched. As a time-scaling algorithm does not adapt the spectral characteristics of the speech segments accordingly, the result will probably be perceived as a distortion.

Incomplete deletions. In the reverse situation to, problems can occur without any of the system components being responsible. When strong coarticulation occurs in the timing reference but not in the original, the time warping path should specify substantial shortening for the corresponding phones from the original. The transitions into and away from these phones in the time-scaled result could then seem too slow and too carefully produced, resulting in a synchronized utterance that appears to be missing some part.

Incompatible substitutions. Some allophones could have different acoustic realizations in the original and the timing reference. If the inherent length of these different realizations differs, the system will produce the acoustic variant from the original with the duration of the timing reference, again leading to perceive a distortion, even in the absence of time warping or time scaling errors. A comparable situation might occur when distortions or noises exist in the original: they could become more prominent after time-scaling.

The above problems are characteristic for the transplantation concept in that they occur without that either the time-warping or the time-scaling procedure can be held responsible for the distortion. When noise is added to the timing reference signal or when the speaker of the timing reference is different from that of the original signal, the distortion problems appear to occur with increased frequency compared to the single speaker clean-speech situation, which is consistent with the hypothesis that problems are related to acoustic-phonetic differences.

4.3. Semi-automatic postsynchronization

Many of the distortions that occur should be easy to identify visually in the time warping path as they will be characterized by long

stretches with either very large or very small derivatives (viz., near horizontal or vertical).

An editor for the warping path was developed [10]. It provides a visual display of the path, and of the utterances $U1$, $U2$ and Ux (Fig. 3). A zoom function can be operated in any of the visual

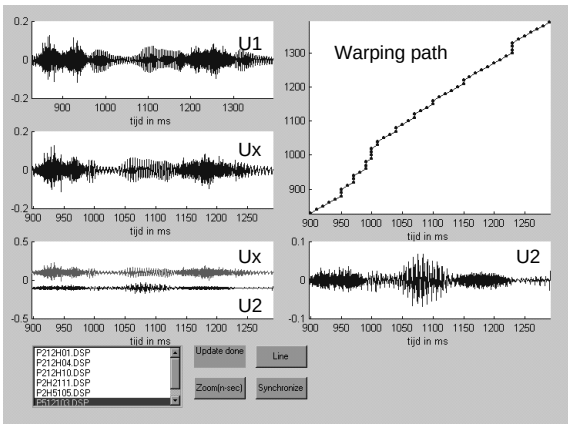

Figure 3: *Illustration of the warping-path editor*

displays and automatically adapts all other displays accordingly.

The utterances can be listened to by clicking on the appropriate waveform plot. There always is a choice to play the full waveform or the displayed portion only.

An edit function in the window of the warping path allows any portion of the path to be modified, after which the warping path itself and Ux can be updated correspondingly.

A few informal tests performed so far seem to indicate that distortions can indeed be easily located and straightforwardly removed with this editor. Thus, the timing transplantation together with the editor could form an effective tool for the semi-automatic correction of lip-sync errors. Further work is needed to confirm this and to verify whether the editor could also be used to make semi-automatic super-karaoke practicable.

5. CONCLUDING DISCUSSION

High quality transplantation of voice characteristics is made possible with DTW and OLA speech modification. With increasing acoustic-phonetic differences between the input utterances, the frequency and severity of audible distortions also increases. In order to make transplantations of voice characteristics between recordings applicable in practice, further work is needed to combat these rejection phenomena.

In some informal tests with the postsynchronization application, an interactive warping path editor was successfully applied to remove the distortions that occurred by smoothing-out discontinuities and abrupt transitions in the path. Thus, it should be interesting to experiment with constraints in the DTW optimization of the path. However, care should be taken in this matter since the occurence of very shallow or steep slopes could correspond to genuine timing differences, such as in the case of breathing pauses, for example.

After its introduction in [1], the idea of prosody transplantation has been adopted by the text-to-speech (TTS) community where it is used to synthesize speech with a prosodic contour that is copied from a natural utterance [11]. While prosody transplantation has become a household term in TTS, we are not aware of rejection phenomena in that context. The main differences with inter-signal prosody transplantation appear to be that, in the TTS case, the prosodic parameters are stylized when they are coded into the prosody generation module of the synthesizer, manual optimization can be used and, in general, carefully prepared speech material is dealt with instead of spontaneous utterances.

In any case, many interesting challenges seem to lie ahead on the roads towards robust high quality inter-signal transplantations of voice characteristics.

6. REFERENCES

[1] W. Verhelst, M. Borger, “Intra-Speaker Transplantation of Speech Characteristics,” proceedings of Eurospeech’91, pp. 1319-1322, 1991.

[2] E. Moulines, F. Charpentier, “Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones,” Speech Communication, vol. 9, nrs. 5,6, pp. 453-467, 1990.

[3] W. Verhelst, D. Van Compernolle and P. Wambacq, “A Unified View On Synchronized Overlap-Add Methods for Prosodic Modification of Speech,” in Proc. of ICSLP 2000, Beijing, pp. II.63-II.66, October 2000.

[4] J. R. Deller Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Chapter 11, Macmillan 1993.

[5] W. Verhelst ”On The Quality of Speech Produced By Impulse Driven Linear Systems”, proceedings of ICASSP-91, pp. 501-504, 1991.

[6] K.H. Lau, “A System for Hybridizing Vocal Performance,” Masters Thesis, University of Miami, Coral Gables, Florida, 2001.

[7] K.H. Lau, “A System for Hybridizing Vocal Performance,” Audio Engineering Society Convention Paper 5625, presented at the 112th AES Convention, Munich, May 10-13, 2002.

[8] W. Verhelst, “Automatic Postsynchronization of Speech Utterances,” in Proc. of Eurospeech 97, Rhodos, pp. 899-902, September 1997.

[9] W. Verhelst, M. Roelands, “An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech,” proceedings of ICASSP-93, vol. II, pp. 554-557, 1993.

[10] J. Vanderwaeren, “Een systeem voor automatische postsynchronizatie in de audio- en video industrie,” Masters Thesis, ESAT-KULeuven, 2001.

[11] B. Van Coile, L. Van Tichelen, A. Vostermans, J.W. Chang, M. Staessens, ”Protran: A prosody transplantation tool for text-to-speech applications”, proceedings of ICSLP-94, pp. 423-426, Yokohama, 1994.